Métodos y técnicas de detección de unidades terminológicas

Mercedes Ramírez Salado
Instituto Universitario de Investigación en Lingüística Aplicada
Universidad de Cádiz
mercedes.ramirez@uca.es

Vanesa Álvarez Torres
Instituto Universitario de Investigación en Lingüística Aplicada
Universidad de Cádiz
vanesa.alvarez@uca.es

Resumen

El trabajo terminográfico requiere una gran precisión metodológica, especialmente en la fase inicial, pues de ello dependerá que las unidades constitutivas de objeto de análisis sean auténticos términos. Para poder seleccionar correctamente estos candidatos de estudio existen diversos métodos, entre los que destacaremos los usados en el marco del proyecto "Comunicación especializada y terminografía: usos terminológicos relacionados con los contenidos y perspectivas actuales de la semántica léxica" y en la tesis doctoral adscrita que se está desarrollando de forma paralela al citado proyecto, titulada "Terminología y lingüística forense: usos terminológicos relacionados con los ámbitos de actuación de la lingüística forense y su interfaz con otras disciplinas".

El principal objetivo de este trabajo es exponer distintos métodos para la extracción de términos y los resultados obtenidos de su puesta en práctica en las investigaciones mencionadas. Desde el punto de vista del proyecto de investigación, se ha trabajado desde una perspectiva que podríamos describir como tradicional y que supone la detección de candidatos a término fundamentada en la documentación y la revisión por expertos, mientras que, en la tesis doctoral adscrita, se ha realizado la detección mediante un método automático basado en técnicas pertenecientes a la lingüística de corpus.

Palabras clave: terminología, terminografía, detección semiautomática de términos, metodología.

1. Introducción

Aunque el nacimiento de la terminología tiene lugar en dominios no lingüísticos, esta disciplina está totalmente integrada en el panorama lingüístico actual, ocupando un lugar de enorme trascendencia debido, sobre todo, a su carácter transdisciplinar. En este sentido, surgen cada vez más trabajos centrados en el estudio de las unidades terminológicas de uno u otro ámbito científico. En nuestro caso, en el seno del Instituto Universitario de Investigación en Lingüística Aplicada (ILA), se está desarrollando un proyecto de investigación de excelencia que se ocupa del estudio de los usos terminológicos propios de la semántica léxica y, de forma paralela, se está realizando una tesis doctoral que recopila y analiza los usos terminológicos pertenecientes a otra rama de la lingüística aplicada: la lingüística forense.

En el presente trabajo, expondremos brevemente en qué consiste cada una de las investigaciones centrando nuestra atención en uno de los principales puntos de encuentro entre ambos estudios, esto es, la metodología del trabajo terminográfico.

Desde el proyecto "Comunicación especializada y terminografía: usos terminológicos relacionados con los contenidos y perspectivas actuales de la semántica léxica" (TERLEX), cuyo origen se sitúa en el estudio de Casas Gómez (2006), se ha trabajado para la recopilación de unidades terminológicas que serían posteriormente objeto de estudio desde el punto de vista de los usos terminológicos, es decir, el objetivo principal es aclarar el sentido en que empleamos los términos de este ámbito y definirlos según una adecuada caracterización de sus distintas acepciones terminológicas. En el marco de este proyecto se empezó por delimitar qué aspectos de los planteamientos tradicionales de la vieja lexicología siguen plenamente vigentes en la

semántica léxica más actual y, para ello, 9 investigadores del proyecto participaron en la sección temática sobre "Nuevos contenidos y perspectivas actuales de la semántica léxica" del *XX Deutscher Hispanistentag* (Universidad de Heidelberg, 2015), encuentro del que se ha derivado la publicación de un volumen monográfico sobre semántica léxica (cf. Casas Gómez/Hummel 2017). Una vez delimitados los contenidos y con el fin de analizar los usos terminológicos propios de este ámbito, llegamos a la conclusión de que la fase de detección y selección de términos juega un papel clave, pues será el punto de partida del resto de la investigación.

En cuanto a la tesis doctoral que se desarrolla de forma paralela, cabe destacar que se adscribe al proyecto de investigación mediante un contrato predoctoral y que, si bien tiene también como objetivo principal el análisis de usos terminológicos, en este caso, los usos pertenecen a otra de las áreas de la lingüística aplicada. La tesis, titulada "Terminología y lingüística forense: usos terminológicos relacionados con los ámbitos de actuación de la lingüística forense y su interfaz con otras disciplinas", versa sobre la recopilación y análisis de las unidades terminológicas vinculadas a la lingüística forense y, al igual que en el proyecto TERLEX, las técnicas para la detección de dichas unidades resultan fundamentales para la correcta evolución de la investigación.

2. Métodos utilizados en el marco de nuestras investigaciones

En esta sección detallaremos las distintas técnicas y métodos empleados en las dos investigaciones en las que centramos este trabajo, pues, como hemos comentado anteriormente, uno de los puntos comunes entre ambos estudios es la importancia que posee la fase de detección de unidades terminológicas. En cualquier trabajo terminográfico, la metodología debe estar sumamente clara para lograr que el resultado final de la investigación sea lo más preciso y correcto posible. Para evitar los problemas derivados de una inadecuada planificación, hemos afrontado una serie de decisiones de carácter metodológico, pero con un objetivo fundamentalmente práctico, esto es, la creación de un léxico especializado en cada uno de los ámbitos que nos ocupan. No obstante, debido a las características que presentaban los dos campos (la semántica léxica y la lingüística forense), no ha sido posible seguir los mismos métodos en este proceso. El principal problema radicó en la poca tradición que tiene la lingüística forense como disciplina, lo que causa una casi inexistencia de términos de este sector consignados en los diccionarios de lingüística, por lo que el método lexicográfico presentaba grandes dificultades, mientras que en el campo de la semántica léxica se encontró un gran número de términos recogidos en diccionarios, lo que permitió fundamentar la detección de unidades en técnicas lexicográficas o, como aquí las hemos denominado, tradicionales. De este modo, en el proyecto TERLEX se han empleado métodos tradicionales que detallaremos en el apartado 2.1., mientras que, en la tesis doctoral adscrita, debido a las características de la terminología de la lingüística forense, se ha empleado un método que combina distintas técnicas de extracción de términos.

2.1. Técnicas tradicionales

En el proyecto "Comunicación especializada y terminografía: usos terminológicos relacionados con los contenidos y perspectivas actuales de la semántica léxica", las unidades de estudio se han seleccionado empleando, fundamentalmente, dos técnicas de corte tradicional. Por una parte, se han seleccionado obras lexicográficas del ámbito de la lingüística que datan de los últimos 50 años y se han revisado sus entradas para constatar cuáles de ellas remitían a contenidos propios de la semántica léxica actual. Para agilizar la labor de búsqueda dentro de los diccionarios, estos se han digitalizado y sometido al reconocimiento óptico de caracteres, lo que ha permitido buscar dentro de los textos con mayor facilidad. El proceso de conversión de las fuentes lexicográficas a formato digital se ha llevado a cabo mediante un escáner cenital de

última generación (Zeutschel OS12002 Advanced Plus), que dispone de unas características específicas para digitalizar documentos en alta resolución y contribuir a la preservación de los mismos. Este equipamiento pertenece al Gabinete de Asesoría lingüística del Instituto Universitario de Investigación en Lingüística Aplicada (ILA) de la Universidad de Cádiz, que cuenta con un Servicio Periférico de Investigación (SPI) de digitalización en alta resolución a través del que se han digitalizado las fuentes lexicográficas y doctrinales para el proyecto TERLEX. Si bien este proceso constituye una fase más en el presente trabajo, creemos oportuno detallar el procedimiento llevado a cabo en la digitalización del material bibliográfico, pues se trata de un proyecto, dirigido por el profesor Casas Gómez, consistente en la creación de un repositorio electrónico de documentación lingüística de la futura biblioteca del ILA[1]. En cuanto al proceso de digitalización, es relevante añadir que todas las obras escaneadas quedan registradas en una tabla de 47 campos con información bibliográfica de las mismas, así como el código del archivo de trabajo (Scanjob), por si necesitamos volver a trabajar sobre el documento en cuestión.

Las obras digitalizadas cuentan con unos metadatos que facilitan la organización y el manejo de los archivos y, de este modo, se guardan en carpetas ordenados alfabéticamente por autores, ya que el software del escáner, Omniscan 12, está configurado para crear automáticamente una carpeta con el nombre del autor del documento digitalizado y, de esta manera, recopilar todos los archivos del mismo autor bajo la misma carpeta. Los metadatos que se introducen al iniciar la digitalización son tres: 1) autor, 2) título y 3) año. Si tomamos como ejemplo la obra *Diccionario de lingüística moderna* de los autores Enrique Alcaraz Varó y Mª Antonia Martínez Linares, los metadatos quedarían representados con el siguiente formato: 1) autor: Alcaraz Varó, E. y Martínez Linares, M. A., 2) título: Diccionario de lingüística moderna y 3) año: 1997. Por tanto, el nombre del archivo resultante sería: Alcaraz Varó, E. y Martínez Linares, M. A._1997_Diccionario de lingüística moderna. Una vez finalizada la digitalización de la obra, realizamos el proceso de reconocimiento óptico de caracteres (OCR) con la aplicación ABBYY FineReader 14, pues, de esta manera, se facilita la búsqueda de términos con sus correspondientes contextos de usos[2].

Tras esta primera fase donde el corpus de vaciado estaba conformado por diccionarios especializados, se logró elaborar un listado que contenía 313 unidades terminológicas pertenecientes a la semántica léxica, listado que se entregó a los miembros del proyecto que funcionaron también como revisores en la siguiente fase. En este caso, el equipo de investigación, conformado por integrantes del grupo de excelencia "Semaínein" del Plan Andaluz de Investigación, garantizaba la fiabilidad de la revisión, debido a su amplia trayectoria investigadora en torno a la semántica en general y la semántica léxica en particular. Esta segunda fase consistió, primeramente, en la revisión pormenorizada de las unidades para corroborar su inserción en nuestro campo de interés y, posteriormente, en la inclusión de nuevas unidades que, sin estar consignadas en el listado hasta el momento, se consideraban propias de la semántica léxica actual. Tras esta última etapa, el listado quedó ampliado a un total de 523

[1] Además de las obras lexicográficas, se han digitalizado fuentes doctrinales sobre los distintos campos actuales de la semántica léxica en el marco del proyecto de excelencia "Comunicación especializada y terminografía: usos terminológicos relacionados con los contenidos y perspectivas actuales de la semántica léxica" (TERLEX), dirigido por Miguel Casas Gómez, que formará parte del repositorio documental de la futura biblioteca del Instituto. Este material bibliográfico se ha clasificado temáticamente en materias y se ha establecido una serie de palabras claves, facilitándose, de este modo, la organización y la búsqueda de los mismos.

[2] La extracción de términos, en un principio en obras lexicográficas, se ha ampliado a fuentes doctrinales y se ha procedido a la creación de un corpus metalingüístico para el análisis de los usos de los términos estudiados en el proyecto TERLEX.

unidades terminológicas que actualmente están siendo analizadas para establecer los usos terminológicos que presentan cada una de ellas.

2.2. Técnica combinada

En la tesis doctoral titulada "Terminología y lingüística forense: usos terminológicos relacionados con los ámbitos de actuación de la lingüística forense y su interfaz con otras disciplinas", se pretende, de manera análoga al proyecto de excelencia en el que se adscribe (TERLEX), recopilar y analizar los usos terminológicos propios de la lingüística forense, disciplina que, si bien está en auge, no cuenta con demasiados trabajos escritos en español y, los que hay, presentan ciertos problemas terminológicos, por lo que consideramos necesario realizar un trabajo de carácter terminológico en esta rama de la lingüística.

Dado que existe un claro punto de encuentro entre el proyecto y la citada tesis, partimos de la idea de mantener la misma metodología de trabajo, pero pronto percibimos que los diccionarios de lingüística no consignan suficientes términos de la lingüística forense, por lo que debemos recurrir a otras técnicas para la detección de unidades terminológicas. De este modo, surge la necesidad de crear un corpus de vaciado distinto a los materiales lexicográficos con los que se contaba hasta el momento, concretamente, se seleccionaron un total de medio centenar de obras que tratan de algún modo la lingüística forense y/o sus campos de actuación, siguiendo las pautas propuestas por Pérez (2002). Así, el corpus de vaciado está formado por textos especializados que se digitalizaron siguiendo las pautas empleadas en el proyecto TERLEX y cuyo formato se adapta a texto plano (txt). Esta necesidad de tratar los textos con los que trabajaríamos desde el ámbito computacional ya la expone, en terminología, Drouin (1997: 47), cuando afirma que "l'information nécessaire à de tels traitements doit inévitablement être encodée par l'humain dans une étape de pré-traitement du corpus".

Una vez obtuvimos el corpus de vaciado en el formato deseado, optamos por el uso de un software específico para la extracción de candidatos a término y evitar así posibles problemas derivados del estudio subjetivo de los textos. Para ello, llevamos a cabo una breve revisión de los principales sistemas empleados para la detección de términos basándonos, sobre todo, en el trabajo de Estopà (1999), que aporta datos de gran significancia, puesto que realiza un estudio detallado de diversos extractores elaborados entre los años 1989-97, y Estopà (2001), que trata los elementos necesarios para la mejor creación de un sistema de extracción automática. Después de comprobar las características de algunos de los sistemas actuales, decidimos trabajar con un sistema reciente y sumamente accesible: TermoStat Web 3.0, una herramienta híbrida basada en conocimiento lingüístico y estadístico desarrollada por Drouin (2003), que se sustenta en técnicas propias de la lingüística de corpus y de la lingüística computacional, calculando y comparando frecuencias de forma automática mediante el contraste de corpus especializados y generales. Así, nuestro corpus en el formato de texto plano es introducido en este sistema que es accesible mediante un sitio web que solo requiere de la creación de un usuario para poder trabajar y que contiene previamente un corpus general, con textos no técnicos, con 30.000.000 de ocurrencias que usa como base de comparación. Tras los distintos cálculos, TermoStat Web 3.0 nos devuelve un listado con unas 6.000 unidades que detecta como candidatos a término. Este listado es sometido a un primer filtrado en el que se eliminan palabras comunes que el sistema proponía como candidatos; se eliminan términos propios de otras disciplinas lingüísticas y no lingüísticas que, aunque aparecieran en el corpus, no son representativas y características de la lingüística forense; se unen términos complejos que el sistema proponía como términos distintos siendo realmente parte de una misma unidad y, finalmente, se incluyen términos utilizados en nuestra área de estudio y que el programa no ha devuelto, ya sea por no aparecer en el corpus de análisis o por obtener una baja puntuación durante el cálculo estadístico. De este modo, el listado se reduce considerablemente,

conformándose un listado provisional con 150 unidades terminológicas que, al igual que en el proyecto de referencia, se somete a una revisión por parte de expertos, en este caso, contamos con los coordinadores del gabinete de Lingüística forense del Instituto Universitario de Investigación en Lingüística Aplicada de la Universidad de Cádiz, que, si bien realizan apreciaciones sobre algunas de las unidades terminológicas, corroboran la fiabilidad de los resultados obtenidos en las fases anteriores.

Debido a la suma de métodos aquí empleados (método lexicográfico, método automático y su correspondiente filtrado y revisión por expertos), hemos denominado al conjunto de procedimientos como *técnica combinada*, puesto que supone la combinación de distintos métodos para lograr extraer las unidades terminológicas.

3. Conclusiones

Se deduce de estudios de este tipo la creciente importancia de las nuevas tecnologías en las investigaciones lingüísticas, dado que constituyen una herramienta de trabajo muy útil que facilitan el desarrollo de la investigación, sin dejar de lado la labor del lingüista como encargado de planificar, controlar y revisar el procedimiento. En nuestro caso, se trata de dos estudios terminológicos que se apoyan en disciplinas auxiliares, como son la documentación, la informática y la lingüística de corpus, pero empleando técnicas distintas en función de las necesidades de cada trabajo.

Por una parte, en el proyecto TERLEX, por la consolidación que ya poseen los términos de la semántica léxica en los diccionarios de lingüística, se ha podido partir de un corpus lexicográfico digitalizado para llegar a una revisión por expertos, que finalizó con la inclusión de nuevos términos y usos terminológicos que, actualmente, mediante herramientas de lingüística de corpus, están siendo constatados y analizados dentro de un corpus especializado compuesto por casi 6.000 trabajos que tratan, de un modo u otro, la semántica léxica.

Por otra parte, en el campo de la lingüística forense, aunque también tenemos el objetivo de analizar los usos terminológicos, no nos ha sido posible seguir el mismo método para la detección de términos, sobre todo, debido a la ausencia de términos de este ámbito en los diccionarios. Sin embargo, este inconveniente ha podido solventarse mediante la confección de un corpus especializado que se sometió a diversos tratamientos y análisis automáticos para extraer los candidatos a término. Finalmente, siguiendo la metodología más tradicional, el listado de unidades terminológicas se sometió a una revisión por expertos para corroborar y completar los resultados emitidos por el sistema, es decir, en este caso hemos empleado la combinación de varias técnicas.

Finalmente, observamos que la relación de la terminología con la documentación y la informática constituye una prueba más del carácter interdisciplinar de los estudios terminológicos, mientras que la aportación de la terminología a la semántica léxica y a la lingüística forense, tras nuestras investigaciones, hace patente el carácter transdisciplinar que posee la terminología.

Referencias bibliográficas

Casas Gómez, M. (2006). ‘‘Modelos representativos de documentación terminográfica y su aplicación a la terminología lingüística’’. *Revista de Lingüística y Lenguas Aplicadas*, 1, 25-36.

Casas Gómez, M. y Hummel, M. (eds.) (2017). *Semántica léxica*, volumen monográfico de *RILCE. Revista de Filología Hispánica*, 33, 3. Pamplona: Universidad de Navarra.

Drouin, P. (1997). “Une méthodologie d'identification automatique des syntagmes terminologiques: l'apport de la description du non-terme”. *Meta*, 42:1, 45-54.

Drouin, P. (2003). “Term extraction using non-technical corpora as a point of leverage”. *Terminology*, 9:1, 99-115.

Estopà Bagot, R. (1999). *Extracció de terminologia: elements per a la construcció d’un SEACUSE (Sistema d’Extracció Automàtica de Candisats a Unitats de Significació Especialitzada)*. Tesis Doctoral, Barcelona: Institut Universitari de Lingüística Aplicada

Estopà Bagot, R. (2001). “Extracción de terminología: elementos para la construcción de un extractor”. *TradTerm*, 7, 225-250.

Pérez Hernández, C. (2002). “Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento”. *Estudios de Lingüística del Español*, 18. Publicado en: <https://ddd.uab.cat/pub/elies/elies_a2002v18/431.html> [Consultado el 15 de marzo de 2018].